

Review

Looking forward: eye-gaze methods in vocabulary development research

Haley Weaver^{1,*}, Jenny Saffran ², and Sudha Arunachalam³

Human multimodal processing abilities have provided researchers with an invaluable set of methods for interrogating language understanding. Even young infants fixate on visual stimuli that match incoming auditory information. Experimental paradigms have harnessed this behavior to demonstrate early language comprehension abilities. Researchers have since adapted these paradigms to address new questions, such as studying individual differences in vocabulary size and structure, identifying which words are learned earlier or later, and assessing language in populations with disabilities. However, fundamental questions persist about the assumptions linking eye gaze with underlying linguistic competence. We aim to articulate these assumptions and outline what we know about whether they are met. By making these issues explicit, we highlight considerations for language development research across different populations.

Detecting language knowledge in infants and toddlers

How do we know what babies know? If only we could just ask them! Despite the inherent challenges in ascertaining knowledge in beings who can't answer questions or push buttons, researchers have developed ingenious ways of interrogating the minds of infants (and of nonhuman animals, who also can't directly respond to our queries about what they know).

Infant visual behavior provides an invaluable window into the infant mind [1]. Early uses of eye-gaze methods with infants focused on basic perceptual abilities, later extending to studies of learning that compared infants' preferences for trained versus novel stimuli (for a brief overview of nonlinguistic eye-gaze work with infants, see Box 1). However, these sorts of forced-choice paradigms left open many questions about the nature of infants' knowledge. In particular, infants' nascent knowledge about their native language remained stubbornly opaque. It cannot be studied in nonhuman animals (unlike, say, visual perception), limiting the informativeness of cross-species comparisons. Additionally, the timeline of language production is quite protracted; what infants say (or sign) vastly understates what they want to say, as any parent of a frustrated toddler can tell you.

This review examines how eye-gaze methods have been developed and used in efforts to address these methodological barriers in the study of early lexical knowledge. After providing an overview of the genesis and main uses of this method, we delve into foundational assumptions that merit closer scrutiny, particularly given the wide range of research questions and populations now being studied with eye-gaze methods.

The evolution of eye-gaze methods for studying early word knowledge

In the late 20th century, two developments converged to provide a new approach to studying early language knowledge. The first was the implementation of multimodal tasks designed to

Highlights

Eye-gaze measures are an invaluable tool for investigating language knowledge in infants and toddlers.

Despite their many strengths, eye-gaze measures rely on linking assumptions connecting observable behavior with underlying linguistic knowledge that merit closer attention.

This review focuses on three key assumptions underpinning the application of eye-gaze methods to studying early word comprehension: (i) infants look at images when they are named; (ii) eye-gaze measures index graded knowledge; (iii) eye-gaze measures can tell us whether individual infants know specific words.

We conclude with some promising directions, as well as ongoing interpretative puzzles. We also explore implications for clinical applications and the assessment of children with language-related disabilities.

¹Department of Speech, Language, and Hearing Sciences, Boston University, Boston, MA 02215, USA

²Department of Psychology, University of Wisconsin-Madison, Madison, WI 53711, USA

³Department of Communicative Sciences and Disorders, New York University, New York, NY 10012, USA

*Correspondence: hjweaver@bu.edu (H. Weaver).

Box 1. Eye gaze in infant research in domains beyond language

Tasks using eye gaze are ubiquitous across the infancy literature. Indeed, a recent analysis of papers published in the journal *Infancy* between 2016 and 2018 found that nearly half of the papers used looking behavior as a dependent variable [2]. Infancy researchers have used gaze behavior to tap into infants' early perceptual and cognitive abilities for over 50 years. The initial development of eye-gaze methods for studying language comprehension by Golinkoff *et al.* [3], was inspired by research on infants' audiovisual perception, demonstrating that infants will preferentially attend to visual stimuli that match the audio stimuli that they are hearing (e.g., specific rhythms, numbers, and vowels [4,5]).

In the visual domain, habituation paradigms have long considered infants' eye movements as an indicator of whether or not an infant has habituated to a stimulus. Early work with these tasks defined thresholds for infants' looking behavior; infants are considered to have habituated if they look markedly less than they did on the initial exposure—for example, if the most recent 3 trials show 50% of the looking observed on the first 3 trials [6]. Using these gaze-based methods, a rich body of work has uncovered impressive abilities ranging from early face perception [7–9] and categorization [10,11] to social and moral development [12,13].

Visual preference and habituation paradigms have a long history of challenges, despite their wide adoption in the developmental literature. In particular, concerns regarding what 'counts' as novel or familiar, and how preferences for different kinds of stimuli develop, have complicated interpretations. In addition, there may be cultural differences in infant looking behavior that complicate interpretation of data drawn from communities beyond the typical **WEIRD population** of study participants [14]. Large-scale global efforts to ameliorate these methodological concerns have been of broad interest in recent years [15–19]. As an example, in the upcoming ManyBabies 5 project, researchers are examining what factors (i.e., age, familiarity, and complexity) influence infants' visual preference for stimuli—thus testing methodological assumptions behind a commonly used paradigm with larger sample sizes and a more culturally diverse population of infants than any individual study can offer [18]. These efforts demonstrate a growing enthusiasm for better understanding factors that influence looking-time measures.

determine whether infants connected auditory and visual events, as indexed by their gaze to visual stimuli that either matched or mismatched an auditory stimulus [4,5]. Building on studies in the nonlinguistic literature, Golinkoff, Hirsh-Pasek, and their colleagues developed a parallel method to investigate whether infants would fixate on the referent of a familiar word after hearing it [3]. Their pioneering study demonstrated that when viewing pairs of objects (images) or events (videos) on the screen, infants would shift their gaze based on linguistic information. The key dependent variable in their **Intermodal Preferential Looking Paradigm (IPLP)** (see Glossary) was **accuracy**: the percentage of time spent looking at the **target** image or scene within a **target window** (a preset period of time following the onset of the target's label).

The second development was the application of remote eye tracking to measure real-time language processing in adults using a method known as the **Visual World Paradigm** [20]. In these tasks, adults' eye-gaze behavior was measured as they heard speech instructing them to engage with objects in their immediate environment, with particular attention to the time course of fixations. Fernald and her colleagues developed a screen-based version of this task to begin characterizing infants' **reaction time** during language comprehension, an index of lexical access [21]. As in the Golinkoff *et al.* IPLP task, infants viewed pairs of objects on the screen; the key difference was the focus on the time course (e.g., reaction time to shift gaze from the distractor object to the target object) as a key dependent variable that might differ across infants, across ages, etc. In later work, Fernald *et al.* named their version of this task **Looking While Listening (LWL)** [22]. Using this paradigm, Fernald and colleagues demonstrated that infants become faster at recognizing words across development, and that the speed and accuracy with which a child recognizes words predict language skills later in childhood [21,23,24]. Figure 1 depicts an example of the setup (Figure 1A) and trial timing (Figure 1B) for a word recognition task using the LWL method. Figure 1C illustrates the time course of infant eye gaze during a typical word recognition study.

Glossary

Accuracy: the proportion of looks to the target image during the target window, divided by the total number of looks to the target image plus the distractor image. In some analyses, looks to neither image are included in the denominator.

Convergent validity: the extent to which two measurements align. The key assumption is that the two measurements theoretically index the same underlying construct and, therefore, should be correlated.

Distractor image: eye-gaze tasks with infants typically include a pair of images. The distractor image is the one that is not named. Features of the distractor image can be manipulated to examine effects on target recognition.

Intermodal Preferential Looking Paradigm (IPLP): the first eye-gaze method for examining word recognition, developed by Golinkoff, Hirsh-Pasek, and colleagues. Infants view a pair of images or videos, and spoken language guides them to fixate on a visual target.

Looking While Listening (LWL): developed by Fernald, Swingle, and colleagues, this method builds on the IPLP paradigm to focus on the time course of infants' fixations on the visual target.

ManyBabies: a global consortium of research labs, ManyBabies (<https://manybabies.org/>) studies bring together researchers using the same stimuli and tasks to generate unusually large and diverse samples for individual infant studies.

Mutual exclusivity bias: the assumption that a new word refers to a new object rather than a familiar object.

Reaction time: for trials where the infant starts out looking at the distractor image, their reaction time is the first frame on which they fixate on the target image.

Reliability: the precision of a particular measurement. Measurements that are reliable should be similar across trials.

Saccades: rapid eye movements that shift visual fixation from one location to another.

Target image: the referent of the spoken word, usually a visual image. In some studies, the target may be related to the spoken word (e.g., sharing its color or other semantic features) rather than being identical to it.

Target window: the analytic window used to ascertain word recognition.

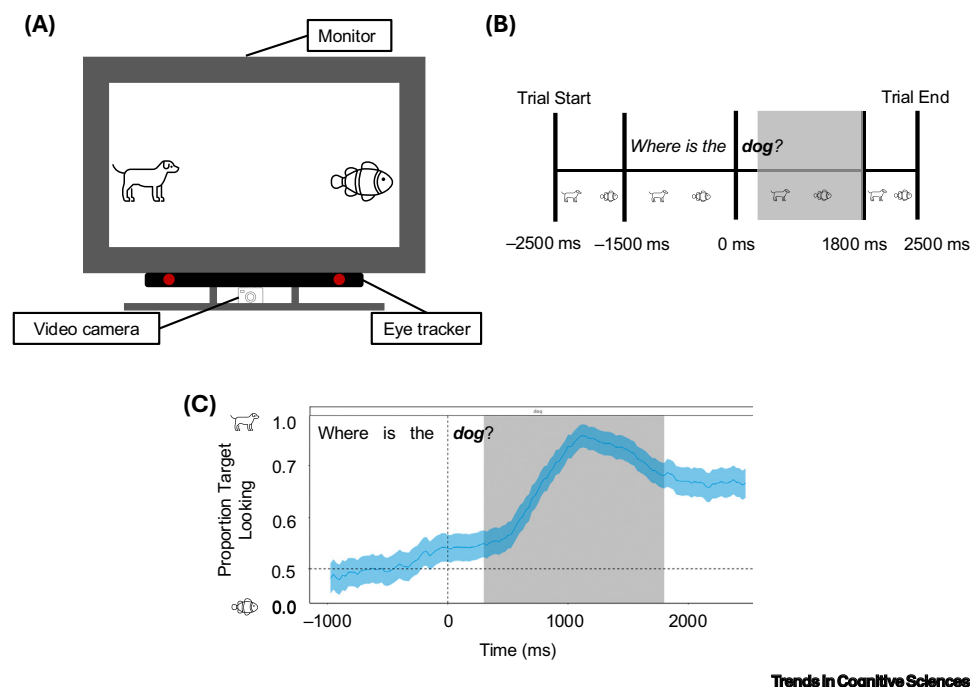


Figure 1. Setup, trial structure, and gaze time course in a word recognition paradigm. Panel (A) It depicts the set-up of a typical eye-gaze word recognition task. The infant would be centered in front of the screen, and would see two images presented concurrently, along with speech directing their attention to one of the images (the target). Panel (B) It depicts the timeline for a representative trial in a word comprehension task. At the start of the trial, the target (dog) and distractor (fish) images are presented in silence. The shaded portions of Panel B and Panel C represent the critical window of analysis after the onset of the target noun ('dog'), as suggested in [22] (300–1800 ms). Panel (C) It provides representative data for trials where the target noun is 'dog'. The data were generated using the PeekBank Visualization Tool and include all available datasets with infants ages 0–24-month-old that tested the word 'dog' [25]. The X-axis depicts time from the onset of the trial; 0 ms represents the onset of the target word 'dog'. The Y-axis depicts the proportion of fixations to the target (looks to the target divided by looks to the target + looks to the distractor, averaged across participants). As shown in Panel C, looking behavior starts out at chance (50%) and then quickly transitions to the target after the onset of the spoken noun.

Typically, the window begins 200–300 ms after the onset of the target word, allowing for a saccade based on the spoken language, and ends 2–3 s later. In predictive processing tasks, this window may occur prior to the onset of the target word.

Validity: the extent to which a measurement indexes the construct of interest. A measurement must be valid to draw conclusions from the results of the study.

Visual World Paradigm: this method was pioneered by Tanenhaus and colleagues to measure real-time word recognition in the context of objects. In a typical study, adult eye gaze is tracked while they listen to speech and interact with a set of objects that are manipulated to test hypotheses about word recognition (e.g., phonological or semantic overlaps).

WEIRD populations: this acronym refers to people from western, educated, industrialized, rich, and democratic societies, who are heavily overrepresented in psychological research.

Trends in Cognitive Sciences

Contemporary applications of eye-gaze methods

While early instantiations of eye-gaze tasks aimed to capture evidence of infants' emerging word comprehension, these paradigms have been adopted broadly in developmental psycholinguistics [26]. Here, we focus on word comprehension studies (leaving aside the rich literature using these methods to study word learning, phonological processing, and other aspects of early language development). One burgeoning area of research has focused on understanding individual differences in the size, composition, and structure of early vocabularies. For example, children who know more words that denote object categories organized by shape (e.g., cup) represent the shape of an object (rather than other features such as color) more strongly than children with fewer of those words [27]. Similarly, lexical connectivity (i.e., vocabularies with more semantic neighbors and/or words with more shared features) has different influences on word recognition across development [28]. Related lines of research have asked whether differences in vocabulary size or structure could be caused by differences in the language input infants are exposed to. Indeed, the type and quantity of language input that infants receive predict their ability to recognize words [29–31]. However, these input characteristics are neither necessary nor sufficient to predict infants' word knowledge [32].

The field has also become increasingly interested in uncovering which specific words infants know and what they may know about those words. For instance, young infants, as a group, understand frequently used food and body part words [33,34] and can accurately apply labels to both specific individuals [35,36] and object categories [37–39]. Infants are also sensitive to phonological and semantic relations between words, as evident from studies manipulating the degree of similarity between targets and distractors [29,40–43]. However, as we will examine later, scholars have increasingly noted methodological limitations in eye-gaze tasks that preclude strong interpretations of children’s knowledge of individual words [44–46], potentially posing challenges for understanding precisely when and what words are learned across development.

Although current uses of eye-gaze tasks largely mirror the original purposes of the task (i.e., to reveal language comprehension prior to production), many challenges remain, particularly as researchers pursue more specific questions about the nature of emergent lexical knowledge and as increasingly interdisciplinary teams attempt to apply these methods to new populations of children. The low task demands of eye-gaze tasks, relative to assessment measures that require pointing or language production, make them attractive as a way to tap vocabulary knowledge in hard-to-test populations, including very young infants [33,35,36] and toddlers with developmental conditions such as autism [47]. However, doing so raises new issues for analysis and interpretation (see Box 2 for a discussion of the use of eye-gaze methods in research with clinical populations).

As the use of these methods continues to be extended beyond their original focus, we risk losing sight of key (implicit) assumptions about what eye-gaze behavior can and cannot reveal about lexical development. In the next section, we lay out three major assumptions underpinning eye-gaze methods as applied to the development of word comprehension, and some of the issues that arise when considering them in the context of contemporary research questions.

Box 2. Using eye-gaze methods with clinical populations

Children with developmental disabilities pose unique challenges for clinical assessments of language. In particular, assessment instruments require children to use expressive language, engage in social interaction, and/or deploy motor abilities which may be limited in clinical populations. Gaze-based tasks, by contrast, have very low task demands. Children only need to have normal or corrected hearing and vision and sufficient postural control to look at a screen, making it a viable paradigm for children with a wide range of abilities. As such, eye-gaze methods have increasingly been used to assess language comprehension in clinical populations [48].

Gaze-based tasks have proven to be sensitive to emerging language comprehension abilities across children with a wide range of developmental disabilities (e.g., autism, cerebral palsy, cochlear implants, Down syndrome, and Fragile X syndrome). This work has demonstrated word recognition prior to caregiver reports and traditional assessments [47,49–53]. Given these benefits, researchers are excited about the opportunity to leverage eye gaze as a clinically relevant measure of receptive vocabulary [54]. These implicit tasks could be immensely useful in identifying how many and which words a given child knows, thereby informing and individualizing intervention approaches.

However, adapting methods from research with typically developing children is not straightforward. Many of the design concerns described in this review may interact differently with attentional or processing skills in clinical populations. For example, children with developmental delays or disorders may have generally slower processing speeds [43] or difficulty disengaging from visual stimuli in response to a prompt [55], thus requiring longer windows of analysis to capture comprehension and complicating comparison with typically developing peers. That is, a child who is slower to identify a target referent than a typically developing peer, but indeed still looks reliably at a referent, must be separable from a child who is unable to reliably identify a referent. By only adopting analytical windows used with typically developing children, these subtle differences in ability may be masked by overt group differences. Relatedly, there is vast within-group heterogeneity in both language development and gaze behavior, making it challenging to disentangle differences in comprehension from differences in task behavior. Further, once methodological issues are resolved, clinically relevant cut-offs for accuracy or reaction time measures will need to be identified. These issues deserve experimental attention so that progress can be made toward integrating eye tracking into valid and reliable receptive language assessment tools.

Assumption no. 1: infants look at images when they are named

The most fundamental assumption underlying these paradigms is that connecting sound and image is essentially mandatory: if an infant understands a word, they will shift their gaze to its referent. This assumption is generally well-founded [3,56]. From birth, infants can fixate on a visual target and launch **saccades** (eye movements) between targets [57,58], and these abilities improve throughout childhood. By 6–9 months of age, well before they begin to produce words, infants' deployment of fixations and saccades reflects their recognition of words. Infants in this age range look preferentially at the referents of common words such as 'mommy', 'milk', and 'banana' when prompted to do so (e.g., 'look at the milk' [33–36]). That said, we may see variability in the degree to which this assumption plays out as a function of task or experimental design factors, and these decisions can impact the conclusions that are drawn.

To successfully connect words with images, children must recognize that the word they are hearing refers to one of the images they are viewing. This is not a trivial task; infants experience a limited set of individual referents and, over time, learn to extend their word meanings to entire categories of referents [37–39]. Moreover, infants' and young children's everyday experiences may not present them with the stereotypical referents for words that are depicted in research studies, and this may be particularly true for more abstract words, such as early routines (e.g., *uh-oh* and *hi*), verbs, and emotion words [59–62].

In another sense, whether, when, and how much infants look at a named referent also depends on what else there is to look at—that is, what the **distractor image** or images are. For instance, a child may look more at an image of a golden retriever than an image of a sock after hearing 'dog', simply because animate objects such as dogs are more interesting to look at than inanimate objects such as socks [63,64]. While some distractor effects on accuracy can be managed statistically (e.g., by adjusting the data from the target window to reflect baseline visual preferences), they can also hinder the interpretability of the data, particularly when reaction time is a dependent variable.

Infants' knowledge about the label for the distractor object might also influence their looking behavior toward the target object in these two-alternative forced-choice tasks. It has been well established that, as early as 17 months of age, young children use their prior word knowledge to constrain their novel word learning [65–68]. This **mutual exclusivity bias** also influences gaze behavior during LWL tasks; infants look more at a named target object when they can correctly reject the distractor as the referent of the label [46]. That is, word recognition may sometimes inflate infants' apparent knowledge about a target label due to their knowledge of the label for the distractor object. This influence of distractor knowledge may be even stronger in yoked experimental designs, where two objects are consistently paired throughout the task. In such designs, a child's looks to a particular target may always be driven by its known distractor, irrespective of the child's knowledge about the target label. By contrast, unyoked designs that vary the distractor items across test trials may assess target words in the presence of both known and unknown distractors, reducing the influence of prior knowledge. Ultimately, the stimulus pairs included in the task influence the extent to which gaze behavior solely reflects knowledge of the target words.

More generally, the specific auditory and visual stimuli included in a given experiment impact measures of word recognition [46,59,63,69]. Infants are better at identifying target words when they are paired with distractor objects that have lower frequency or a later age of acquisition [46,69–71]. Infants also respond to features of objects that partially match their lexical representations; infants (as well as older children and adults) will look more at a red car than a blue car

when hearing, ‘Can you find the strawberry?’ [72,73]. Cross-language stimuli—such as sentences that switch languages prior to the target word—demonstrate how manipulating the auditory information in these tasks can provide insight into the organization of multilingual lexicons [74]. It is evident that performance in these eye-gaze tasks reflects not just the recognition of a single target word, but also the semantic networks and language systems within which those target words are embedded [28,75]. The fact that eye-gaze word comprehension methods tap into broader aspects of the lexicon is both a feature and a bug. This interconnectedness facilitates the development of studies focused on semantic structure and the nature of the lexicon, but it may also hinder the interpretation of studies designed to examine infants’ knowledge about individual words.

Assumption no. 2: eye-gaze measures index graded knowledge

The second assumption underlying these paradigms is that they illuminate graded vocabulary knowledge. That is, gaze behavior is measured continuously and is aggregated across time, yielding continuous measures of speed and accuracy. Often, these numerical differences are interpreted as reflecting overall differences in lexical knowledge, for example, infants with larger versus smaller vocabularies. These numerical differences may also reflect the robustness of individual word representations, for example, an infant who knows more about a specific word that is being tested may show different patterns of looking than an infant who knows less about that word.

A burgeoning body of research demonstrates that individual differences in gaze-based measures are associated with other validated vocabulary measures, providing evidence for **convergent validity**. Indeed, children who are faster to recognize familiar words (and to anticipate upcoming words [Box 3]) have larger vocabularies, as reported by their caregivers or as indexed by their own performance on a standardized assessment [31,88–91]. Similarly, gaze behavior reflects the structure and composition of the vocabulary, such that children with more densely connected vocabularies demonstrate more adultlike word recognition [92–94]. Given this wealth of evidence,

Box 3. Predictive language processing

Eye tracking has been used extensively to assess broader language processing abilities in children. One such paradigm, initially developed with adult participants, is often referred to as verb-mediated prediction [76]. The results suggest that adults predict upcoming linguistic information after hearing a semantically constraining sentential context (e.g., looking at a cake rather than a shoe after hearing the verb ‘eat’). Infants and children also demonstrate this impressive ability, using phonological information [77,78], semantic relationships [79–81], morphosyntactic structure [82,83], and information about the speaker such as their sex and age [84] to guide their looking behavior before hearing a target word.

Prediction and incremental language processing, as evidenced by anticipatory looks to a target, have been implicated as key abilities related to language learning. Indeed, predictive processing abilities are correlated with measures of native language attainment [79,85,86]. More broadly, this literature has demonstrated incredible continuity with adult language processing; as children gain more advanced linguistic knowledge, they leverage their lexical networks to initiate anticipatory looks to phonological and semantic competitors.

Stimuli in predictive processing tasks tend to be heavily constrained—the relationship between predictor words (such as verbs) and targets (such as nouns) are tightly controlled, and thus many design-specific assumptions are made explicit in these paradigms. However, timing challenges remain. What is the appropriate window of analysis following the semantic context over which one should assess prediction? For example, to observe verb-mediated predictions about upcoming nouns, do children have to hear the whole verb or can they make predictions based on partial information [87]? Is a boost in target looking sufficient to provide evidence of predictive processing, or do looks need to precede the onset of the target? Although it is well established that lexical processing efficiency improves across development, unknowns about the time course of particular gaze behaviors make it challenging to implement this paradigm in new populations or using nonstandard stimuli (i.e., using stimuli other than targets entailed by the semantic or phonological context).

it is perhaps unsurprising that we often assume that graded differences in word recognition reflect real differences in vocabulary size and structure.

There is also an expansive literature aiming to capture individual differences in the robustness of word knowledge: differences in speed/accuracy index differences in how deeply a child knows the tested words. Studies that intentionally manipulate semantic relatedness demonstrate that gaze behaviors are influenced by the extent to which a child represents aspects of meaning beyond the visual referent [29,40,95,96]. Other work aims to characterize the precise nature of early word meaning representations [27,38,39,42]. For instance, one study found that older infants look significantly less at incorrect but semantically related target objects than younger infants, suggesting that older children have more refined word representations [42]. In such studies, looking provides a behavioral marker of how quickly, easily, and correctly a child can retrieve a word's meaning.

At the same time, slower or less accurate looking could reflect slower processing, which might be driven by a language or cognitive delay rather than anything specific about the items being used in an experiment [49,97]. Moreover, there are a host of biological, rather than cognitive, explanations for slowed reaction times. For example, one child might blink more or saccade more slowly than another. Critically, these alternative interpretations might not necessarily mean that children's knowledge of the words is less robust, but rather that they require more time to demonstrate their knowledge. Slower saccades would also likely result in less target looking because children who take longer to reach the target have less time to fixate on it before the trial moves on. Paradoxically, faster looking may also result in short durations if the time window for analysis is too long because children will eventually look away from a target and toward the distractor. These individual differences may be inherently interesting for understanding cognitive processing, but they make it challenging to disentangle the speed and accuracy of eye gaze from the robustness of word knowledge within a single experimental paradigm.

A related challenge is that word recognition effects in eye-gaze tasks are often quite small, with proportions of target looking after the onset of the noun on the order of 55–60% (vs. chance, 50%) [30,46,92,98]. It is currently an open question whether these small, graded differences in infants' gaze behavior reflect variation in the strength of children's word knowledge or whether other processes may account for such differences. That is, do children who fixate on the target referent more quickly and for a longer proportion of time actually have more robust representations of the word's meaning, the phonological form of the label, the connection between the two, or something else entirely? Efforts to investigate individual differences in word recognition while holding other factors constant (e.g., age and diagnostic status) or including additional tasks designed to tease out other factors may help us better discern the roots of these sources of variability (see the Looking forward section).

Assumption no. 3: eye-gaze measures reveal whether individual infants know specific words

Eye-gaze word comprehension methods were initially developed to provide information about overall vocabulary knowledge. Early versions of these tasks included just 1–2 trials for each target word, which were appropriate for estimating broad vocabulary information but insufficient to draw conclusions about specific word knowledge [99]. However, some more recent studies do attempt to use eye-gaze data as an index of word knowledge at the child level [46,50,54,61,92,98,100,101].

There are several challenges that arise when using eye-gaze data to determine whether a given child knows a given word. First, multiple trials per word and multiple image referents are needed

to ensure that children have adequate opportunity to display their knowledge about individual words [92,102]. However, there is an inherent tradeoff between the number of test items per word and the number of different words tested, particularly given the brevity of an infant study. Experiments that probe individual word knowledge using multiple trials per word are limited in the number of different words they can test, hindering generalizability. Moreover, word repetitions are vulnerable to the design challenges raised previously: Should word pairs be yoked or unyoked? How might children learn across the task, such that trials purportedly measuring word recognition actually reflect word learning strategies (e.g., using mutual exclusivity to refine guesses about label-object mappings as the task proceeds)?

Second, determining whether a child knows a specific word requires identifying a cutoff point for defining 'knowledge'. But what is the correct cutoff? Does fixating on the target at any level above chance indicate that the child knows the target word? This criterion was adopted in the recently developed National Institutes of Health (NIH) Baby Toolbox [54]. Other studies use more conservative numbers, for example, as a 15% increase from a baseline to a test period [61,103–105], or scores above a 95% confidence interval around chance [46]. One study aimed to identify the best threshold for determining whether a child knew an individual word by testing several options (ranging between 5 and 25% change in looking to the target from baseline to test) [105]. However, the 25% threshold determined from this work has not been universally adopted, and it remains unclear how broadly it applies (e.g., beyond the particular stimuli and ages tested).

The difficulty with testing assumptions about what counts as 'knowing' a word is that we lack ground truth about which words a given infant knows. Older children and adults can point, press buttons, and answer questions. While these responses are also not ground truth, at least they are (relatively) clearly operationalized and make intuitive sense. With infant eye-gaze data, binary decisions about whether or not a baby knows a word are based on a continuous variable that, as noted above, is influenced by many factors that may or may not be indicative of actual word knowledge. Parent report, though valuable, may also not reflect ground truth at the level of individual words (particularly for comprehension, which is more difficult to observe), challenging interpretation [106–108]. Indeed, stimulus factors (e.g., age of acquisition and mutual exclusivity), caregiver factors (e.g., confidence in word-level comprehension ratings), and learner characteristics (e.g., age and vocabulary size) impact the alignment between eye-gaze and caregiver report measures [44,46].

Looking forward: considerations for researchers

Although there is much experimental work yet to be done, the foregoing review suggests several design principles that researchers may consider in guiding future research. First, it is crucial to identify the extent to which we can use eye gaze to measure individual word knowledge with **reliability** and **validity**, particularly if the goal is to assess individual children's vocabularies (as in many clinical applications). Recent efforts to increase statistical power in looking-time tasks provide key insights into issues of reliably capturing individual differences, such as the value of increasing the number of trials [109]. Measuring validity is less straightforward, especially when considering constructs where ground truth is inaccessible. One possibility is to compare eye-gaze measures of word knowledge to convergent measures with the same participants, such as pointing [110] and caregiver report [44,46]. For example, a study using a dataset drawn from multiple word-learning studies compared looking and pointing across a large sample of preschoolers ($N = 914$), revealing convergent evidence for these measures [110]. This strategy could be applied to examining key design decisions such as window sizes and yoked versus varied stimulus pairings.

Many of the studies reviewed above aim to identify specific parameters that influence the interpretation of children's eye gaze in these tasks. This kind of research is crucial, but it requires large

sample sizes, which can be difficult to achieve within a single lab. **ManyBabies**, a global consortium designed to answer research questions with a sample size in the thousands, offers an excellent model [15–18]. Another promising avenue is Peekbank [25], an eye-gaze data repository that allows researchers to ask questions (e.g., effects of age, stimulus differences, etc.) with more data than could be collected in a single lab.

Another major challenge is the sheer number of variables that differ across studies (e.g., window size, item selection, and operationalization of eye-gaze measures). There are also a number of variables that can be extracted from eye-gaze data that are less widely used (e.g., shape of the fixation curve over time, fixation density, longest fixation, and first fixation durations), and so we do not know how useful they are in revealing more nuanced aspects of lexical knowledge. Some combination of these variables may yield the best estimate of word knowledge; for example, a longer window may be appropriate for children who saccade often or for stimuli that are more visually engaging.

We suspect that researchers (ourselves included) examine such variables in an exploratory fashion but do not report them when they are not predictive in analytic models. Preregistering and publishing a wider set of variables could be highly informative about which measures are appropriate for which types of research questions. Machine learning might also offer a way to address these challenges; with a sufficiently large dataset, gaze patterns, rather than single variables, could be identified that are associated with another measure (e.g., children's pointing or parent report).

Another approach is to add additional tasks to tease apart the roots of individual differences. For example, eye-gaze word recognition tasks that require the child to disengage when they are fixating on the incorrect object (e.g., if they hear, 'Where's the shoe?' and are not already looking at the shoe, they need to shift their gaze to the shoe in order to demonstrate word recognition). Individual differences in visual attention, as well as potential group differences (Box 2), could impact performance for reasons unrelated to word recognition. To address this concern, one study included a separate visual task to measure disengagement [55]. Performance on the disengagement task accounted for half of the variance in word recognition behavior in a sample of autistic children. This study provides a model for including control tasks that can help account for factors that are potentially distorting the effects researchers are trying to study (see [63] for a similar strategy with a different outcome).

Concluding remarks

In articulating the assumptions underlying eye-gaze word recognition paradigms, we wanted to first call attention to their existence. Second, we endeavored to highlight some of the interpretive challenges they raise (see [Outstanding questions](#)). This is especially important for research using language assessments to determine areas in which children might need additional support or intervention. More generally, it is critical to be clear about the questions these methods are being used to address and confident that we are interpreting eye-gaze measures as appropriate to the research question. We hope that the incredible progress made in tapping infants' underlying knowledge and skills only continues to grow with the advent of new technologies, analytic approaches, and theoretical developments.

Acknowledgments

Preparation of this manuscript was supported by the University of Wisconsin–Madison College of Letters & Science, the New York University School of Culture, Education, and Human Development, and grant R01 HD105313 to J.S. The ideas in this review have benefited from discussions with many members of each of our lab groups, to whom we are very grateful.

Outstanding questions

Is it possible to establish 'ground truth' for whether or not a child knows a word? We have suggested that it is not, but we welcome challenges to this perspective.

In the absence of ground truth, how will we know if eye-gaze measures are valid? To determine if these tasks measure what we think they measure will require triangulation across a range of known, validated measures of children's language. Each measure has limitations, but taken together, we can be more confident that we understand the relations among these validated measures and eye-gaze measures.

Is it possible to know enough about what these tasks actually measure to answer clinically relevant research questions? A crucial step forward in answering this question will be rigorous research on individual differences and clinical population-level patterns to identify whether we can draw valid conclusions from these measures.

Can these research tasks be scaled up into tools that can be applied across clinical and educational settings without sacrificing validity? The stakes for using these measures in a clinical or educational context are high; children's qualification for services or targets of intervention may depend on getting these answers correct.

Declaration of interests

The authors declare no competing interests.

References

- Aslin, R.N. (2007) What's in a look? *Dev. Sci.* 10, 48–53
- LoBue, V. et al. (2020) The importance of using multiple outcome measures in infant research. *Infancy* 25, 420–437
- Golinkoff, R.M. et al. (1987) The eyes have it: lexical and syntactic comprehension in a new paradigm. *J. Child Lang.* 14, 23–45
- Spelke, E.S. (1979) Perceiving bimodally specified events in infancy. *Dev. Psychol.* 15, 626–636
- Kuhl, P.K. and Meltzoff, A.N. (1982) The bimodal perception of speech in infancy. *Science* 218, 1138–1141
- DeLoache, J.S. (1976) Rate of habituation and visual memory in infants. *Child Dev.* 47, 145–154
- Quinn, P.C. and Tanaka, J.W. (2009) Infants' processing of featural and configural information in the upper and lower halves of the face. *Infancy* 14, 474–487
- Pascalis, O. et al. (2002) Is face processing species-specific during the first year of life? *Science* 296, 1321–1323
- Gaither, S.E. et al. (2012) Biracial and monoracial infant own-race face perception: an eye-tracking study. *Dev. Sci.* 15, 775–782
- Younger, B.A. and Cohen, L.B. (1983) Infant perception of correlations among attributes. *Child Dev.* 54, 858–867
- Elmas, P.D. and Quinn, P.C. (1994) Studies on the formation of perceptually based basic-level categories in young infants. *Child Dev.* 65, 903–917
- Hamlin, J.K. and Wynn, K. (2011) Young infants prefer prosocial to antisocial others. *Cogn. Dev.* 26, 30–39
- Kinzler, K.D. et al. (2007) The native language of social cognition. *Proc. Natl. Acad. Sci. U. S. A.* 104, 12577–12580
- Beckner, A.G. et al. (2024) Examining infants' visual paired comparison performance in the US and rural Malawi. *Dev. Sci.* 27, e13439
- ManyBabies Consortium (2020) Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Adv. Methods Pract. Psychol. Sci.* 3, 24–52
- Byers-Heinlein, K. et al. (2020) Building a collaborative psychological science: lessons learned from ManyBabies 1. *Can. Psychol.* 61, 349–363
- Frank, M.C. et al. (2017) A collaborative approach to infant research: promoting reproducibility, best practices, and theory-building. *Infancy* 22, 421–435
- Kosie, J.E. et al. (2023) ManyBabies 5: a large-scale investigation of the proposed shift from familiarity preference to novelty preference in infant looking time. *OSF* https://doi.org/10.31234/osf.io/ck3vd_v1
- Lucca, K. et al. (2025) Infants' social evaluation of helpers and hinderers: a large-scale, multi-lab, coordinated replication study. *Dev. Sci.* 28, e13581
- Tanenhaus, M.K. et al. (1995) Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634
- Fernald, A. et al. (1998) Rapid gains in speed of verbal processing by infants in the second year. *Psychol. Sci.* 9, 228–231
- Fernald, A. et al. (2008) Looking while listening: using eye movements to monitor spoken language comprehension by infants and young children. In *Developmental Psycholinguistics: Online Methods Children's Language Processing* (Vol. 44) (Seidenberg, P. et al., eds), pp. 97–135, John Benjamins
- Fernald, A. et al. (2006) Picking up speed in understanding: speech processing efficiency and vocabulary growth across the second year. *Dev. Psychol.* 42, 98–116
- Marchman, V.A. and Fernald, A. (2008) Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Dev. Sci.* 11, F9–16
- Zettersten, M. et al. (2023) Peekbank: an open, large-scale repository for developmental eye-tracking data of children's word recognition. *Behav. Res. Methods* 55, 2485–2500
- Golinkoff, R.M. et al. (2013) Twenty-five years using the intermodal preferential looking paradigm to study language acquisition: what have we learned? *Perspect. Psychol. Sci.* 8, 316–339
- Perry, L.K. and Saffran, J.R. (2017) Is a pink cow still a cow? Individual differences in toddlers' vocabulary knowledge and lexical representations. *Cogn. Sci.* 41, 1090–1105
- Borovsky, A. (2022) Drivers of lexical processing and implications for early learning. *Annu. Rev. Dev. Psychol.* 4, 21–40
- Bergelson, E. and Aslin, R.N. (2017) Nature and origins of the lexicon in 6-month-olds. *Proc. Natl. Acad. Sci. U. S. A.* 114, 12916–12921
- Homer, J. et al. (2025) The home language environment predicts individual differences in language comprehension at 9 months of age. *Dev. Psychol.* 61, 1888–1903
- Weisleder, A. and Fernald, A. (2013) Talking to children matters: early language experience strengthens processing and builds vocabulary. *Psychol. Sci.* 24, 2143–2152
- Foushee, R. and Srinivasan, M. (2024) Infants who are rarely spoken to nevertheless understand many words. *Proc. Natl. Acad. Sci. U. S. A.* 121, e2311425121
- Bergelson, E. and Swingle, D. (2012) At 6–9 months, human infants know the meanings of many common nouns. *Proc. Natl. Acad. Sci. U. S. A.* 109, 3253–3258
- Bergelson, E. and Swingle, D. (2015) Early word comprehension in infants: replication and extension. *Lang. Learn. Dev.* 11, 369–380
- Tincoff, R. and Jusczyk, P.W. (1999) Some beginnings of word comprehension in 6-month-olds. *Psychol. Sci.* 10, 172–175
- Campbell, J. and Hall, D.G. (2022) The scope of infants' early object word extensions. *Cognition* 228, 105210
- Garrison, H. et al. (2020) Familiarity plays a small role in noun comprehension at 12–18 months. *Infancy* 25, 458–477
- Kucker, S.C. et al. (2023) Margarita glasses and high heels: how attention to shape, age, and vocabulary impacts children's recognition of typical and atypical exemplars. *Child Dev.* 94, 603–616
- Weaver, H. et al. (2024) Becoming word meaning experts: infants' processing of familiar words in the context of typical and atypical exemplars. *Child Dev.* 95, e352–e372
- Arias-Trejo, N. and Plunkett, K. (2010) The effects of perceptual similarity and category membership on early word-referent identification. *J. Exp. Child Psychol.* 105, 63–80
- Arias-Trejo, N. et al. (2025) The effects of phonological and semantic similarity on early referent identification. *J. Exp. Child Psychol.* 249, 106112
- Bergelson, E. and Aslin, R.N. (2017) Semantic specificity in one-year-olds' word comprehension. *Lang. Learn. Dev.* 13, 481–501
- Ellis Weismer, S. et al. (2016) Lexical processing in toddlers with ASD: does weak central coherence play a role? *J. Autism Dev. Disord.* 46, 3755–3769
- López Pérez, M. et al. (2025) Infants' knowledge of individual words: investigating links between parent report and looking time. *Infancy* 30, e12641
- Kucker, S.C. and Lorenz, M.G. (2025) Expanding the footnote: how minor methodological decisions have impactful consequences for developmental results. *Front. Dev. Psychol.* 3, 1629396
- Weaver, H. and Saffran, J.R. (2026) Interrogating early word knowledge: factors that influence the alignment between caregiver-report and experimental measures. *Dev. Sci.* 29, e70088
- Venker, C.E. et al. (2016) Brief report: early lexical comprehension in young children with ASD: comparing eye-gaze methodology and parent report. *J. Autism Dev. Disord.* 46, 2260–2266
- Venker, C.E. and Kover, S.T. (2015) An open conversation on using eye-gaze methods in studies of neurodevelopmental disorders. *J. Speech Lang. Hear. Res.* 58, 1719–1732
- Horvath, S. and Arunachalam, S. (2023) Assessing receptive verb knowledge in late talkers and autistic children: advances and cautionary tales. *J. Neurodev. Disord.* 15, 44
- Singh, S.J. et al. (2025) Eye-tracking as a measure of receptive vocabulary in non-verbal children with cerebral palsy. *PLoS One* 20, e0325183

51. Brady, N.C. *et al.* (2014) Eye tracking as a measure of receptive vocabulary in children with autism spectrum disorders. *Augment. Altern. Commun.* 30, 147–159
52. Chita-Tegmark, M. *et al.* (2015) Eye-tracking measurements of language processing: developmental differences in children at high risk for ASD. *J. Autism Dev. Disord.* 45, 3327–3338
53. Plesa Skwerer, D. *et al.* (2016) Comparing methods for assessing receptive language skills in minimally verbal children and adolescents with autism spectrum disorders. *Autism* 20, 591–604
54. Novack, M.A. *et al.* (2025) Development and validation of the NIH Baby Toolbox® language measures. *Infant Behav. Dev.* 80, 102121
55. Venker, C.E. (2017) Spoken word recognition in children with autism spectrum disorder: the role of visual disengagement. *Autism* 21, 821–829
56. Trueswell, J.C. (2015) Studying language processing using eye movements. In *Research Methods in Child Language* (Hoff, E., ed.), pp. 177–190, John Wiley & Sons
57. Luna, B. *et al.* (2008) Development of eye-movement control. *Brain Cogn.* 68, 293–308
58. van Renswoude, D.R. *et al.* (2020) Looking (for) patterns: similarities and differences between infant and adult free scene-viewing patterns. *J. Eye Mov. Res.* 13, 1–20
59. Casey, K. *et al.* (2023) Moving beyond “nouns in the lab”: using naturalistic data to understand why infants’ first words include uh-oh and hi. *Dev. Psychol.* 59, 2162–2173
60. Bergelson, E. and Swingle, D. (2013) The acquisition of abstract words by young infants. *Cognition* 127, 391–397
61. Valteau, M.J. *et al.* (2018) An eye-tracking study of receptive verb knowledge in toddlers. *J. Speech Lang. Hear. Res.* 61, 2917–2933
62. Ogren, M. *et al.* (2025) How do emotion words impact children’s emotion learning? *Dev. Psychol.* 61, 641–648
63. Pomper, R. and Saffran, J.R. (2019) Familiar object salience affects novel word learning. *Child Dev.* 90, e246–e262
64. Venker, C.E. *et al.* (2021) Competing perceptual salience in a visual word recognition task differentially affects children with and without autism spectrum disorder. *Autism Res.* 14, 1147–1162
65. Markman, E.M. and Wachtel, G.F. (1988) Children’s use of mutual exclusivity to constrain the meanings of words. *Cogn. Psychol.* 20, 121–157
66. Halberda, J. (2003) The development of a word-learning strategy. *Cognition* 87, B23–B34
67. Byers-Heinlein, K. and Werker, J.F. (2009) Monolingual, bilingual, trilingual: infants’ language experience influences the development of a word-learning heuristic. *Dev. Sci.* 12, 815–823
68. Lewis, M. *et al.* (2020) The role of developmental change and linguistic experience in the mutual exclusivity effect. *Cognition* 198, 104191
69. Potter, C.E. and Lew-Williams, C. (2024) Frequent vs. infrequent words shape toddlers’ real-time sentence comprehension. *J. Child Lang.* 51, 1478–1488
70. Kartushina, N. and Mayor, J. (2019) Word knowledge in six- to nine-month-old Norwegian infants? Not without additional frequency cues. *R. Soc. Open Sci.* 6, 180711
71. Steil, J.N. *et al.* (2021) No evidence of robust noun–referent associations in German-learning 6- to 14-month-olds. *Front. Psychol.* 12, 718742
72. Forbes, S.H. and Plunkett, K. (2019) Infants show early comprehension of basic color words. *Dev. Psychol.* 55, 240–249
73. Johnson, E.K. *et al.* (2011) Toddlers’ language-mediated visual search: they need not have the words for it. *Q. J. Exp. Psychol.* 64, 1672–1682
74. Byers-Heinlein, K. *et al.* (2017) Bilingual infants control their languages as they listen. *Proc. Natl. Acad. Sci. U. S. A.* 114, 9032–9037
75. Wojcik, E.H. (2022) The map trap: why and how word learning research should move beyond mapping. *WIREs Cogn. Sci.* 13, e1596
76. Altmann, G.T.M. and Kamide, Y. (1999) Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* 73, 247–264
77. Mahr, T. *et al.* (2015) Anticipatory coarticulation facilitates word recognition in toddlers. *Cognition* 142, 345–350
78. Pomper, R. *et al.* (2021) Coarticulation facilitates lexical processing for toddlers with autism. *Cognition* 214, 104799
79. Borovsky, A. *et al.* (2012) Knowing a lot for one’s age: vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *J. Exp. Child Psychol.* 112, 417–436
80. Mani, N. *et al.* (2016) “Proactive” in many ways: developmental evidence for a dynamic pluralistic approach to prediction. *Q. J. Exp. Psychol.* 69, 2189–2201
81. Abu-Zhaya, R. *et al.* (2022) Do children use multi-word information in real-time sentence comprehension? *Cogn. Sci.* 46, e13111
82. Günther, T. *et al.* (2023) The development of predictive gender processing strategies during noun phrase decoding: an eye-tracking study with German 5- to 10-year-old children. *J. Speech Lang. Hear. Res.* 66, 3907–3924
83. Lew-Williams, C. and Fernald, A. (2007) Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychol. Sci.* 18, 193–198
84. Borovsky, A. and Creel, S. (2014) Children and adults integrate talker and verb information in online processing. *Dev. Psychol.* 50, 1600–1613
85. Mani, N. and Huettig, F. (2012) Prediction during language processing is a piece of cake—But only for skilled producers. *J. Exp. Psychol.* 38, 843–847
86. Prescott, K.E. *et al.* (2022) Predictive language processing in young autistic children. *Autism Res.* 15, 892–903
87. Breitlefeld, E. and Saffran, J.R. (2025) Environmental context scaffolds children’s semantic representation of novel words. *Cognition* 262, 106162
88. Lany, J. *et al.* (2018) Infants’ lexical processing efficiency is related to vocabulary size by one year of age. *Infancy* 23, 342–366
89. Koenig, A. *et al.* (2020) Lexical processing of nouns and verbs at 36 months of age predicts concurrent and later vocabulary and school readiness. *J. Cogn. Dev.* 21, 670–689
90. Marchman, V.A. *et al.* (2010) How vocabulary size in two languages relates to efficiency in spoken word recognition by young Spanish–English bilinguals. *J. Child Lang.* 37, 817–840
91. Peter, M.S. *et al.* (2019) Does speed of processing or vocabulary size predict later language growth in toddlers? *Cogn. Psychol.* 115, 101238
92. Borovsky, A. *et al.* (2016) Semantic structure in vocabulary knowledge interacts with lexical and sentence processing in infancy. *Child Dev.* 87, 1893–1908
93. Borovsky, A. and Peters, R.E. (2019) Vocabulary size and structure affects real-time lexical recognition in 18-month-olds. *PLoS One* 14, e0219290
94. Borovsky, A. (2022) Lexico-semantic structure in vocabulary and its links to lexical processing in toddlerhood and language outcomes at age three. *Dev. Psychol.* 58, 607–630
95. Styles, S.J. and Plunkett, K. (2009) How do infants build a semantic system? *Lang. Cogn.* 1, 1–24
96. Arias-Trejo, N. and Plunkett, K. (2013) What’s in a link: associative and taxonomic priming effects in the infant lexicon. *Cognition* 128, 214–227
97. Marchman, V.A. *et al.* (2019) Predictors of early vocabulary growth in children born preterm and full term: a study of processing speed and medical complications. *Child Neuropsychol.* 25, 943–963
98. Houston-Price, C. *et al.* (2007) Discrepancy between parental reports of infants’ receptive vocabulary and infants’ behaviour in a preferential looking task. *J. Child Lang.* 34, 701–724
99. Reznick, J.S. and Goldfield, B.A. (1992) Rapid change in lexical development in comprehension and production. *Dev. Psychol.* 28, 406–413
100. Goldfield, B.A. *et al.* (2016) Understanding and assessing word comprehension. *Appl. Psycholinguist.* 37, 529–549
101. Weaver, H. and Saffran, J.R. (2025) Object functions and words reexamined: toddlers’ recognition of function depends on object type. *J. Cogn. Dev.* 26, 816–837
102. Byers-Heinlein, K. *et al.* (2022) Six solutions for more reliable infant research. *Infant Child Dev.* 31, e2296

103. Ertaş, S. *et al.* (2026) Maternal underestimations and overestimations of their infants' word comprehension: effects on mothers' verbal input and infants' receptive vocabulary. *J. Child Lang.* 53, 132–154
104. Killing, S.E.A. and Bishop, D.V.M. (2008) Move it! Visual feedback enhances validity of preferential looking as a measure of individual differences in vocabulary in toddlers. *Dev. Sci.* 11, 525–530
105. Reznick, J.S. (1990) Visual preference as a test of infant word comprehension. *Appl. Psycholinguist.* 11, 145–166
106. Arunachalam, S. *et al.* (2022) Consistency and inconsistency in caregiver reporting of vocabulary. *Lang. Learn. Dev.* 18, 81–96
107. Robinson, B.F. and Mervis, C.B. (1999) Comparing productive vocabulary measures from the CDI and a systematic diary study. *J. Child Lang.* 26, 177–185
108. Lorang, E. *et al.* (2023) Parent certainty and consistency when completing vocabulary checklists in young autistic children. *J. Speech Lang. Hear. Res.* 66, 2750–2765
109. DeBolt, M.C. *et al.* (2020) Robust data and power in infant research: a case study of the effect of number of infants and number of trials in visual preference procedures. *Infancy* 25, 393–419
110. Creel, S.C. (2025) Connecting the tots: strong looking–pointing correlations in preschoolers' word learning and implications for continuity in language development. *Child Dev.* 96, 87–103